Coupling information extraction and data mining for ontology learning in PARMENIDES

Spiliopoulou, M ; Rinaldi, Fabio ; Black, B ; Zarri, G P ; Mueller, R M ; Brunzel, M ; Theodoulidis, B ; Orphanos, G ; Hess, M ; Dowdall, J ; McNaught, J ; King, M ; Persidis, A ; Bernard, L

Abstract: Strategic decision making, especially in the areas of business intelligence and competitive intelligence, requires the acquisition of decision-relevant information pieces like market trends, fusions and company values. This information is extracted by pre-processing and querying multiple sources, combining and condensing the findings. It is characteristic that the extraction process is resource intensive and has to be performed regularly and quite frequently. In the research project PARMENIDES, we are developing methods that establish ontologies over an application domain, annotate documents with the ontology components and identify the entities in them, so that we can decompose business into conventional queries towards entities and XML-annotated texts.

Coupling Information Extraction and Data Mining for Ontology Learning in PARMENIDES

Myra Spiliopoulou[§], **Fabio Rinaldi**[*], **William J. Black**[†], **Gian Piero Zarri**[‡],
Roland M. Mueller[§], Marko Brunzel[§], Babis Theodoulidis[†], Giorgos Orphanos[**],
Michael Hess[*], James Dowdall[*], John McNaught[†], Maghi King[¶],
Andreas Persidis[‖], Luc Bernard[†]

Abstract

Strategic decision making, especially in the areas of business intelligence and competitive intelligence, requires the acquisition of decision-relevant information pieces like market trends, fusions and company values. This information is extracted by pre-processing and querying multiple sources, combining and condensing the findings. It is characteristic that the extraction process is resource intensive and has to be performed regularly and quite frequently. In the research project PARMENIDES, we are developing methods that establish ontologies over an application domain, annotate documents with the ontology components and identify the entities in them, so that we can decompose business into conventional queries towards entities and XML-annotated texts.

1 Introduction

Parmenides is a project in the area of knowledge extraction and management, funded by the IST programme of the 5th Framework for the period 08/2002-01/2005. Its primary aim is the realisation of an ontology-driven systematic approach for integrating the entire process of information gathering, processing and analysis. The project develops novel techniques for the purpose of semi-automatical establishment of domain-specific ontologies, automatic detection and extraction of events in textual data, integration of the extracted information assets in a document warehouse and for temporal knowledge discovery tasks upon the integrated information assets.

In Parmenides, the extraction of knowledge from texts is aimed at (a) the establishment of ontologies which reflect the universe of discourse, (b) the semantic annotation of documents with the concepts, entities and events depicted in the ontologies, (c) the preservation of these semantics in a document warehouse and (d) the acquisition of information and of mining patterns from the document warehouse to support decision making (Orphanos & Tsalidis, 2003). The interplay among these tasks is depicted in Figure 1.

Parmenides exploits two complementary methodologies for the establishment of ontologies over a document corpus. The *Template-oriented methodology* has its origins in natural language processing (and in particular, Information Extraction) and is based upon rule templates that express the structure and semantics of documents in order to enrich them with annotations at different levels of detail (Rinaldi et al., 2003a). The *KDD-oriented methodology* has its origins in data mining and aims at the discovery of concepts and relationships that characterise the document corpus and are relevant for the universe of discourse.

[*]Institute of Computational Linguistics, University of Zurich, Switzerland; [†]Centre for Research in Information Management, UMIST, Manchester, UK; [‡]CNRS, Paris, France; [§]University of Magdeburg, Germany; [¶]TIM/ISSCO, University of Geneva, Switzerland; [‖]Biovista, Athens, Greece; [**]Neurosoft, Athens, Greece;

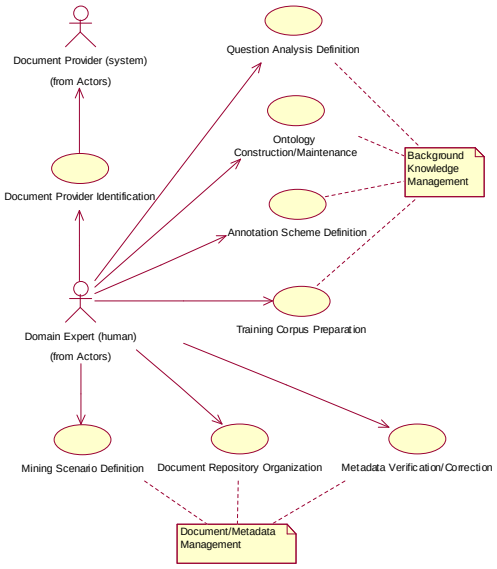

Figure 1: The tasks of Parmenides (Orphanos & Tsalidis, 2003), printed with permission of the PARMENIDES consortium

In section 2 we illustrate the format of the annotations adopted in the PARMENIDES framework (both document-level annotations and conceptual annotations). The next section (3) describes the Information Extraction tools used to create the annotations and the basic Ontology components. Section 4 presents the data mining approach used to propose Ontology Extensions.

The Parmenides system will be demonstrated on four case studies, supplied by the user partners of the consortium: Unilever is a multinational consumer goods manufacturing organization; they contribute to Parmenides two case studies on knowledge management. Biovista is a corporate intelligence organisation in the biotechnology sector; they contribute to Parmenides a case study on market analysis. The IT department of the Greek Ministry of Defence is a governmental organisation with expertise in crisis management; they contribute to Parmenides a case study on text categorisation. Because of space constraints, this paper describes only one such case study in detail (the Biovista case, in section 5).

We conclude the paper with a short discussion about the need for semi-automated Ontology Extension in relation to the wealth of manually built Ontologies being proposed in recent years (section 6).

2 The Annotation Framework

The PARMENIDES project aims at enriching documents with a series of annotations at different levels of complexity (Rinaldi et al., 2003a). The linguistic tools add to documents XML-based annotations according to a specification previously agreed upon by all partners (Rinaldi et al., 2003b). The KDD tools work upon such enriched documents and propose new concepts and classified instances, which are inserted into a complex Knowledge Base built upon the well-known Knowledge Representation Framework NKRL (Zarri, 2003).

Parmenides aims at using consolidated Information Extraction techniques, such as Named Entity Extraction, and therefore this work builds upon well-known approaches, such as the Named Entity annotation scheme from MUC7 (Chinchor, 1997). Other sources that have been considered include the GENIA tagset (GENIA, 2003), TEI (TEI Consortium, 2003) and the GDA tagset (Kôiti). Crucially, attention is also paid to temporal annotations, with the aim of using

extracted temporal information for detection of trends (using Data Mining techniques). Therefore we have investigated all the recently developed approaches to such a problem, and have decided for the adoption of the TERQAS tagset (Ingria & Pustejovsky, 2002; Pustejovsky et al., 2002). The domain of interests (e.g. Biotechnology) are also expected to be terminology-rich and therefore require proper treatment of terminology.

The set of Parmenides annotations is organized into three levels:

- **Structural Annotations**
- **Lexical Annotations**
- **Semantic Annotations**

Structural annotations are used to define the physical structure of the document, it's organization into head and body, into sections, paragraphs and sentences. Lexical annotations identify lexical units that have some relevance for the Parmenides project. Semantic annotations are meant to represent the propositional content of the document (the "meaning"). While structural annotations apply to large text spans, lexical annotations apply to smaller text spans (sub-sentence) and semantic annotations are not directly associated to a specific text span, however, they are linked to text units by co-referential identifiers. All annotations are required to have an unique ID and thus are individually addressable, this allows semantic annotations to point to the lexical annotations to which they correspond. Semantic Annotations themselves are given a unique ID, and therefore can be elements of more complex annotations.

The structure of the documents is marked using an intuitively appropriate scheme based on the TEI recomendations (TEI Consortium, 2003). Broadly speaking, structural annotations are concerned with the organization of documents into sub-units, such as section, title, paragraphs and sentences.

Lexical Annotations are used to mark any text unit (smaller than a sentence), which can be of interest in Parmenides. They include (but are not limited to): Named Entities in the classical MUC sense, new domain-specific Named Entities, Terms, Temporal Expressions, Events. When visualizating the set of Lexical Tags in a given annotated document, clicking on specific tags displays the attribute values, as shown in figure 2.

The relations that exist between lexical entities are expressed through the semantic annotations. So lexically identified people can be linked to their organisation and job title, if this information is contained in the document.

Although the XML-based annotations described so far in this section provide a rich framework for complex document annotations, no reasoning and inferential capability are associated with them. In PARMENIDES, these tasks are then entrusted to NKRL (Narrative Knowledge Representation Language), see Zarri (2003).

NKRL provides a standard, language independent description for the semantic content of narrative documents, in which information content consists of the description of events that relate the real or intended behaviour of some actors.[1] These actors try to attain a specific result, experience particular situations, manipulate some (concrete or abstract) materials, send or receive messages, buy, sell, deliver etc. All the NKRL knowledge representation tools are structured into four connected components:

The descriptive component concerns the tools used to produce the formal representations, called (NKRL) templates, of some general narrative classes of real world events, like moving a generic object, formulate a need, starting a company, obtained by abstraction/generalisation from sets of concrete, elementary narrative events. Templates are inserted into an inheritance hierarchy (a tree) that is called H_TEMP (hierarchy of templates).

The factual component provides the formal representation of the different, possible elementary events characterised, at least implicitly, by precise spatial and temporal coordinates under the form of instances of the templates of the descriptive component. These formal representations are

[1] The term event is taken here in its more general meaning, covering also strictly related notions like fact, action, state, situation etc.

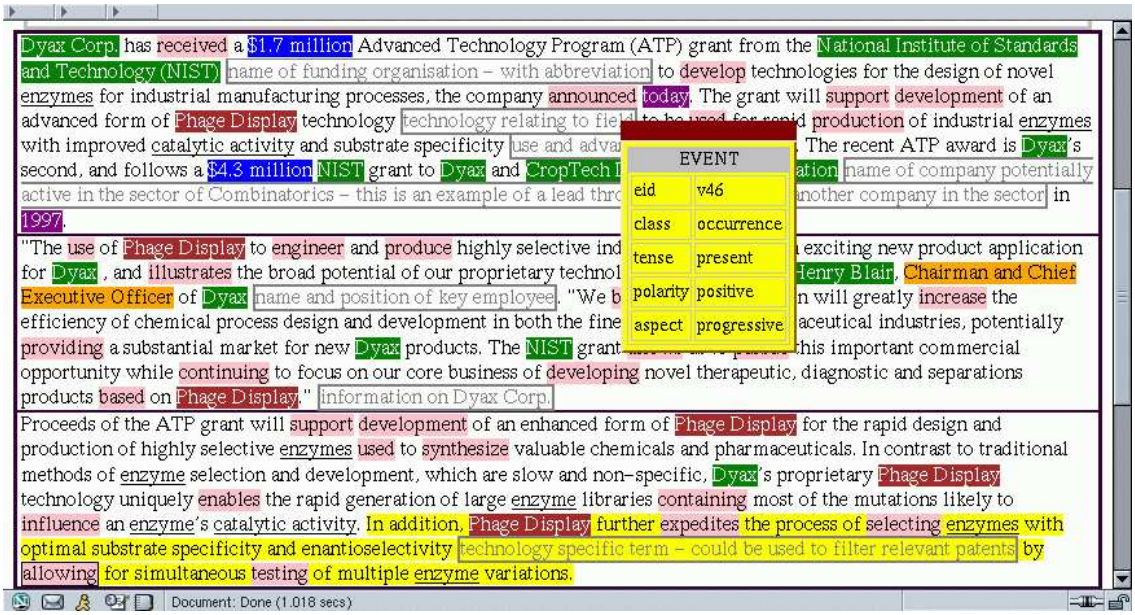

Figure 2: Visualization of Lexical Annotations and their attributes

called (NKRL) predicative occurrences. A predicative occurrence is then the NKRL representation of elementary events like Mr. Smith has fired Mr. Brown.[2]

The definitional component concerns the formal representation of the general notions like human_being, taxi_ (the general class referring to all the possible taxis, not a specific cab), etc. that must be represented for taking into account the events proper to a specific application domain. Their NKRL representations are called concepts , and correspond quite well to the concepts of the usual, formal ontologies of terms. NKRL concepts are inserted into a generalisation/ specialisation directed graph structure, called H_CLASS(es).

The enumerative component concerns the formal representation of the instances (concrete examples) of the general notions (concepts) pertaining to the definitional component; the NKRL formal representations of such instances take the name of individuals. Therefore, individuals are created by instantiating (some of) the properties of the concepts of the definitional component. Individuals are characterised by the fact of being countable (enumerable), of being associated with a spatio-temporal dimension, and of possessing unique conceptual labels (smith_, general_motors): two individuals associated with the same NKRL description but having different labels will be different individuals.

The frames of the definitional and enumerative components are tripartite structures (symbolic label-attribute-value). The descriptive and factual components, however, are characterised by the association of quadruples connecting together the symbolic name of the template/occurence, a predicate and the arguments of the predicate introduced by named relations, the roles. The quadruples have in common the name and predicate components. If we denote then with Li the generic symbolic label identifying a given template/occurrence, with Pj the predicate used (like MOVE, PRODUCE, RECEIVE etc.), with Rk the generic role (slot, case, like SUBJ(ect), OBJ(ect), SOURCE, DEST(ination)) and with Ak the corresponding argument (concepts, individuals, or associations of concepts or individuals), the NKRL data structures for the descriptive and factual components have the following general format:

```
(Li (Pj (R1 A1) (R2 A2) & (Rn An)))
```

[2]Note that the usual ontological languages like description logics (see Buchheit et al. (1993)) do not consider any sort of descriptive or factual structure, which constitute then the main conceptual innovation introduced by NKRL.

We can then say that, in NKRL, the intrinsic properties of concepts and individuals are described as frame-like structures, and that the mutual relationships which can be detected between those concepts or individuals when describing real-world events or classes of events are represented as case grammar-like structures.

Templates are inserted into the H_TEMP(lates) hierarchy, where each node represents a template object, producing a taxonomy of events. This enlarges the traditional interpretation of ontologies where only taxonomies of concepts are taken into consideration. Analogously, all the NKRL concepts are inserted into the H_CLASS(es) generalisation/specialisation hierarchy. At the difference of H_TEMP, which is simply a tree, H_CLASS admits in general multiple inheritance and is, in formal terms, a lattice or DAG, Directed Acyclic Graph.

Individuals (enumerative component), and predicative occurrences (factual component), are linked as well, in a way, with the H_CLASS and H_TEMP hierarchies, where they appear as the leaves of particular concepts and templates. As instances of concepts, individuals share the same basic format (frames) of these last ones; analogously, occurrences are characterised by the same case grammar format of templates. The main reason for keeping the enumerative and factual components separate from the definitional and descriptive ones is linked with the very different epistemological status of, e.g., concepts vs. individuals. Other Knowledge Representation tools used in NKRL for, e.g., representing temporal data, are described in Zarri (1998). The richness and variety of the knowledge representation paradigms used by NKRL - compared with the standard taxonomic (description logic) one used in DAML+OIL, OWL etc. - allows the implementation and use of a variety of reasoning and inference mechanisms neatly more general and powerful than the usual "rule languages" used in the traditional, ontological approach. We will only mention here the possibility of implementing, in NKRL, rules of the "hypothesis" type (automatic construction of causal explanations), of the "transformation" type (allowing to find semantically similar answers also in the absence, in a knowledge base, of the information originally searched for), of powerful (positive and negative) filtering strategies, of case based reasoning (CBR) procedures, etc. Information on these topics can be found, e.g., in Zarri (2003).

3 The Linguistic Pipeline Proposing Basic Ontology Components

Documents are assumed to be gathered from a variety of sources, and thus will present different formats. The first step of processing is going to be a conversion from the source-specific document format to the agreed Parmenides format. This conversion is based on a set of source-specific wrappers (Kushmerick et al., 1997), which transforms the original document into XML structural annotations, as described in section 2. The next step of processing involves addition of basic linguistic information: documents are tokenized, morphologically analyzed and tagged. At this stage sentence boundaries are also detected. This phase completes the creation of the structural annotation, going down to the lowest levels: the sentence and the token.

A Named Entity Extractor is then used to detect persons, organizations, locations and numerical amounts. Together with a terminology extraction tool based on the well-known C/NC algorithm (Frantzi & Ananiadou, 1999) this module creates the basic Lexical Annotations. An example can be seen in figure 2. The Parmenides temporal text mining architecture uses the CAFETIERE (Black et al., 2003) formalism to identify "basic semantic elements" from texts. CAFETIERE stands for "Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations".

The product of the analysis is a set of conceptual annotations as described in section 2. Unlike a 'classical' information extraction (IE) application, the annotations are linked to the classes and instances of an application-oriented ontology, and again, unlike classical IE, provision is made for (human) domain analysts to correct and extend the analyses of the rule-based system prior to the scenario mining step of analysis, to improve its accuracy. Since ultimately, the goal is the discovery of trends in the coincidence of event types, the units to be extracted are occurrences (or

facts). These occurrences are classified relative to a hierarchy of event classes (the NKRL (Zarri, 2003) H_TEMP).

In addition to classification, the temporal grounding of the *event* as indicated by verb tense and aspect, and by temporal adverbials are extracted as features of a lexical annotation. The representation of the occurrence needs the arguments (subject, object, etc.) of the verb (or event-denoting noun) to be identified, to complete a template instance, one of the classes of conceptual annotation supported by the system. The arguments themselves are either named *individual entities* or objects denoted by *terms* in the domain under analysis (identified via the ontology of entities – the H_CLASS in NKRL).

To be extracted during text analysis, some basic semantic elements need to be instances of concepts already in the domain ontology, although others are discovered during analysis. All events whose instances are to be annotated must be in the ontology, but for other elements, the class can be determined heuristically from contextual clues. Not all proper names need to be known to the system prior to analysis, because following the state of the MUC art, it is possible to classify names accurately from their textual occurrence and context. Similarly, not all unnamed entities need to be antecedently known to the system: common noun phrases can be analysed syntactically, or alternatively, annotations can be confined to those for which statistical evidence suggests domain termhood.

The analysis goes through several phases: tokenization, part-of-speech tagging, ontology lookup, and finally rule-based analysis. The tokenizing and tagging modules follow standard techniques. Ontology lookup is currently done using an off-line index from natural language words and phrases to classes, but a direct access to NKRL is currently under development. Rule-based analysis is required for the creation of all lexical annotations above the token level, and all conceptual annotations. Items found in the ontology lookup phase must be confirmed by rules, which may specify contextual constraints that will disambiguate when the same string can name or describe different objects. For example, if "New York" is preceded by a capitalized word or two and a comma, and not followed by a comma or "and" and another capitalized word, it almost certainly names the state rather than the city.

The rule-based analysis formalism is essentially similar to that reported in Black et al. (1997), but enhanced to give various extensions to its expressive power, and now based on a compiled FST implementation. Briefly, phrases and their constituents are described by a set of attribute-value pairs; both negation and disjunction of values are supported; attributes range over orthographic, morpho-syntactic and semantic/conceptual properties; attributes are used as in HPSG-like linguistic formalisms both to constrain and to construct representations by means of feature unification (through Prolog-like named variables); rules are *context-sensitive*, creating annotations only if specified elements are found in the left and/or right context of the phrase; there is a mechanism to identify longer-distance relationships such as anaphoric co-reference. Examples of rules are (1) and (2).

(1)
```
[syn=NP, sem=ORG, sector=EDU, loc=_LOC] =>
 \ [token="University"],
   [token="of"],
   [sem=LOC, token=_LOC] / ;
```

(2)
```
[syn=NNP, sem=PERSON] =>
   [sem=title]{1,2}
 \ [orth=capitalized],
   [orth=upperinitial]?,
   [orth=capitalized] / ;
```

The annotation being constructed is described on the first line of rule (1), by the features `syn`, `sem`, `sector` and `loc`. The first three of these features are ascribed in the rule, but the feature `loc` takes its value from the variable `_LOC`, which *shares* with the other instance which is the value of the `token` feature of the last word in the phrase. (Variables are recognizable to the system by having an initial underscore.) The symbols `\` and `/` mark the boundary between the phrase's

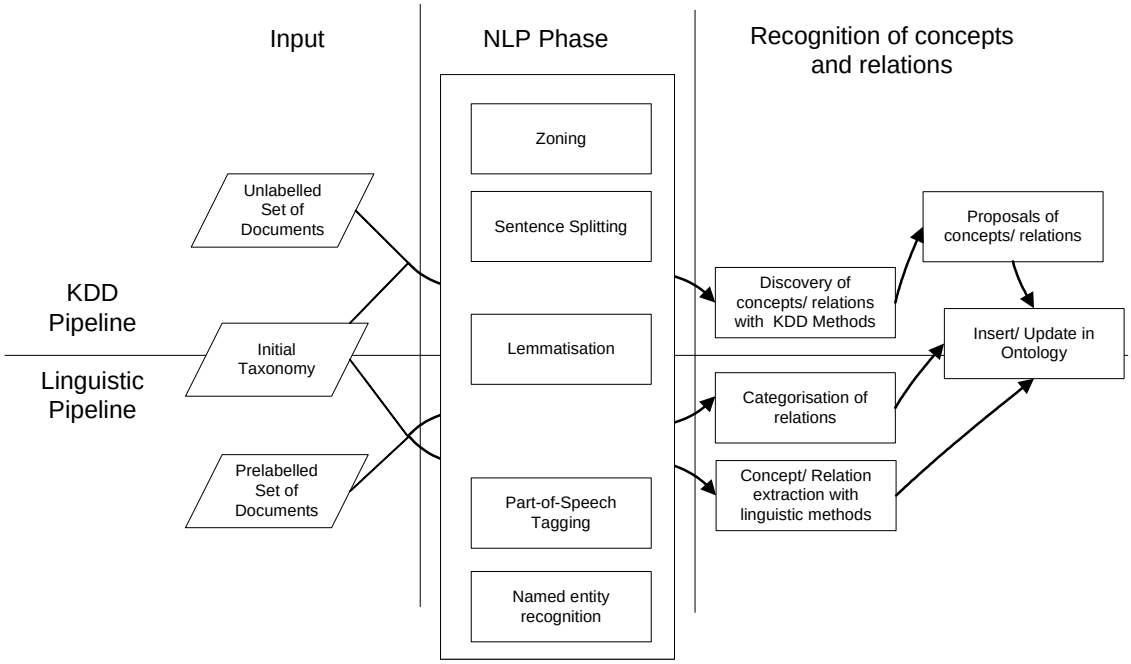

Figure 3: The process of ontology establishment with the two methodologies, printed with permission of the PARMENIDES consortium

constituents and its left and right contexts respectively. In (1) there are no contextual constraints, but in (2) the capitalized words with optional middle initial have to be preceded by a title for the phrase to be considered the name of a person.

There are numerous alternatives to Cafetière, e.g. Cunningham (1999) (which is discussed in relation to an earlier version of Cafetière in Pastra et al. (2002)), Wonsever & Minel (2001) and Xu & Krieger (2003), as well as commercial tools like NLP++. However, in addition to having the expressive power to identify all the types of element on which conceptual indexing and mining can proceed, the system described here is closely integrated with the state of the art annotation framework described above, with annotation tools conformant with the scheme, and in current work, with ontologies and term discovery tools.

4 The KDD Pipeline Proposing Ontology Extensions

The process of ontology establishment in PARMENIDES is depicted in Figure 3. Two complementary methodologies for the establishment of ontologies over a document corpus are adopted and integrated.

The template-oriented methodology aims at (a) the identification and semantic characterisation of multi-term concepts, (b) the discovery and semantic characterisation of relations among concepts, (c) the depiction of entities and (d) the resolution of co-references inside the document corpus. First results are reported in Rinaldi et al. (2003b) and Rinaldi et al. (2003c).

The KDD-oriented methodology aims at the alignment of the taxonomy of the universe of discourse with the document corpora associated with this universe. To this purpose, an existing taxonomy of terms is upgraded into an ontology, using a document corpus.

The two methodologies complement each other:

- The KDD-oriented methodology is intended to annotate large text units like sentences. Non-annotated text units are per se irrelevant to the universe of discourse, so that a finer-grain semantic annotation with tools adhering to the template-oriented methodology can

be avoided. For the remaining text units, the template-oriented methodology undertakes an in-depth semantic characterisation of content, including individual words and multi-word concepts. It also covers the identification of entities and events.

- The KDD-oriented methodology discovers potentially relevant concepts and concept combinations. Then, the template-oriented methodology categorises concept combinations into pre-defined types of relations, so that they can be inserted properly into the application's ontology.

As figure 1 shows, the knowledge extraction process involves at least one human expert, who aligns knowledge extraction to the business objectives and incorporates background knowledge into the Parmenides modules. Parmenides foresees four distinct expert roles:

- The *Domain Expert* possesses deep background knowledge of the application domain and is familiar with the business objectives. This person supervises the establishment of ontologies and defines "scenaria" for information acquisition and knowledge discovery.
- The *NLP Expert* is an expert in natural language processing and responsible for the preparatory NLP steps of text annotation and analysis.
- The *End Users* interact with the Parmenides system to formulate queries and executed data mining scenaria in order to fulfill their information needs.
- The *System Administrator* is responsible for maintenance activities upon the Parmenides system.

The KDD-oriented methodology aims at the alignment of the taxonomy of the universe of discourse with the document corpora associated with this universe. To this purpose, an existing taxonomy of terms is upgraded into an ontology upon to the document corpus:

1. The documents of the corpus are split into sentences, which are mapped into terms of the taxonomy.
2. The vectors of all text units are clustered on similarity.
3. A *quality monitor* selects those clusters, for which labels can be proposed to the domain expert. The selection is based on a combination of criteria on cluster homogeneity, cluster size and relative statistics of the dominant and the non-dominant terms inside each cluster.
4. For each of the selected clusters, the dominant terms are presented and a label is proposed to the domain expert. The domain expert may reject or approve a cluster; in the latter case, the proposed label may be modified.
5. The text units of the remaining clusters undergo the clustering process again, until no further clusters can be selected by the quality monitor.
6. All approved labels are used to annotate the text units of the corresponding clusters.
7. The collection of approved clusters is used to identify similar text units in unknown documents and annotate them automatically.

Approved labels are (i) added in the taxonomy as new concepts or relations among concepts and (ii) used to semantically annotate the text units of the documents. Through the first activity, the taxonomy is upgraded into an ontology that reflects the document corpus. Through the second activity, the document corpus is annotated semantically. The base technology originates from the German research project DIAsDEM (funded by the German Research Foundation DFG under grants SP 572/1, SP 572/3) (Graubitz et al., 2001; Winkler & Spiliopoulou, 2002): DIAsDEM uses clustering techniques and a quality monitor to derive the semantic labels of sentences in an application-specific corpus.

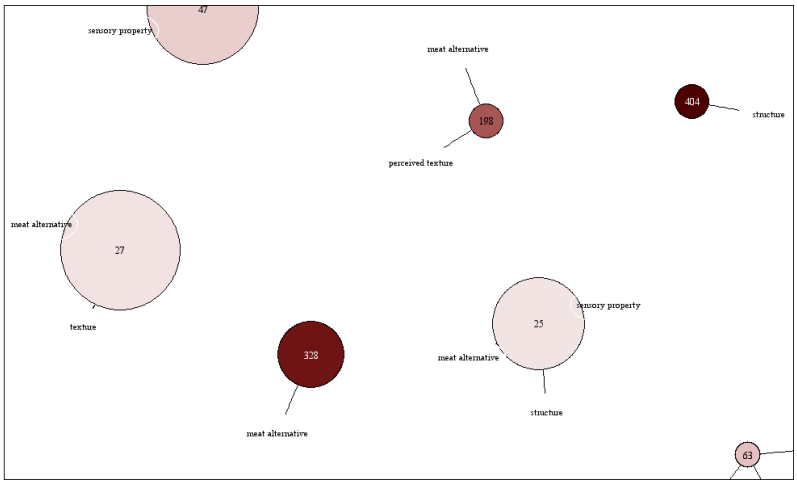

Figure 4: Visualization of a Clustering Result

The ontology established over a text corpus and the annotated text corpus, as stored in the Parmenides document warehouse, are made available to the Domain Expert and the End Users. Their information needs are addressed through the formulation of *questions* and *data mining scenaria*. Briefly, a question is a collection of database queries, while a data mining scenario is a pipeline of KDD tasks to solve a business problem like the prediction of important events or the assessment of the preferences of customers. Parmenides anticipates data mining scenaria for trend analysis and for the temporal evolution of patterns, including at least association rules and frequent sequences.

At the core of the temporal data mining process lays the notion of *event*. An event is a particular type of entity, adorned with a timestamp and with semantic properties (like name) and associated with the document, in which it was traced, and, optionally, with further entities of the application domain. For example, a "fusion"-event is associated with at least three entities: the two (or more) companies that have merged, and the new company that has thus emerged.

Currently, temporal data mining in Parmenides is at a preparatory stage. This includes (i) the conceptual modelling of data mining scenaria according to KD standards, (ii) the specification of event types for the Parmenides case studies, (iii) the development of temporal mining techniques to be applied upon semantically annotated texts and extracted events.

5 An Example Case Study

We have applied the KDD-oriented methodology on one of the four Parmenides case studies, supplied by the user partner Biovista. Biovista is a corporate intelligence organization in the biotechnology sector. At the core of their products are the company's analysts and research staff with expertise in biotechnology, business development, IT, finance and intellectual property issues. Their case study for Parmenides is in the area of market analysis and involves a corpus of business news documents. In the reported case study, we have used 4611 documents.

The goal of applying the KDD-oriented methodology upon the Biovista case documents was to expand their initial, preliminary ontology[3] with new concept combinations. These combinations can be inserted in the taxonomy of the ontology as new concepts or become relations connecting existing concepts. In this goal, the KDD-oriented methodology is intended to discover combinations that were not in the ontology before *and* are frequent in the corpus. Although it is tempting to evaluate this methodology against the expectations of human experts, this would be inappropri-

[3]The ontology is confidential.

ate: The human expert may expect that a concept combination should be found by the software, whilst this combination is too rare in the corpus; or the software may find a concept combination which the expert already knew of but had decided not to include in the corpus for any arbitrary reason. In both cases, we would be evaluating the KDD-methodology against background knowledge that is not available to it. We are currently working on methods allowing an objective evaluation of the findings of the KDD-methodology. For this study, we are disclosing two of the concept combinations that were found.

The first step in our Data Mining process is the specification of the feature space. The preliminary taxonomy provided by the domain expert is aligned to the domain corpus: The taxonomy tree is extended in a way reflecting the concept occurrence frequencies of the corpus, thereby merging sibling concepts which occur rarely and ignoring concepts which do no occur at all. Through this ontology pruning, the original feature space can be reduced to a size specified by the user. In our analysis, we have specified that the feature space of originally 316 dimensions should be reduced to frequent ones, with up to 150 dimensions.

The input to the subsequent clustering process are "text units". Currently, we define as a text unit a sentence. From a total of 142919 sentences in the corpus, we retained those 10920 which contain a least 2 features of the feature space. We have grouped them upon the feature space using the K-Means Clustering Algorithm.

As part of the KDD-methodology, the clusters built in this way are presented to the domain expert as a map and as a bar chart (cf. fig. 4 and 5). The domain expert is responsible for approving a cluster or not, whereby our software marks the "good" clusters according to the following criteria:

- Containment of a non-negligible subset of the text units in the archive: This criterion is motivated by the fact that the derived cluster label is intended to tag the text units in the cluster, and hence should be frequent enough to be useful in search.
- Homogeneity: The cluster label should reflect the content of all cluster members.
- Small number of features describing the cluster: These features constitute the proposed cluster label, which should be short and memorisable.

In Figure 4, we show our visualization of the clustering result, intended to help the user assess the extend, to which a cluster meets the criteria: The circles are clusters, the distance between circles is the 2-dimensional mapping of the cluster-to-cluster distance, the diameter of a cluster is the average intra-cluster distance. The color reflects the cardinality of the cluster; darker clusters are larger. Therefore small, dark circles surrounded by empty space denote good clusters.

We assess the quality of a cluster by considering:

- Feature purity: A cluster is good, if it contains very frequent and (possibly very rare) descriptors, but no descriptors of medium frequency.
- Average intra-cluster distance: A cluster is good, if the average distance from the centroid is small.
- Silhouette: A cluster is good, if its members are closer to each other than to members of other clusters.

Currently, the domain expert can violate the suggestions of the software and reject a "good" cluster or approve a cluster that the software does not consider as "good". All sentences belonging to an approved cluster are labelled with the frequent features of this cluster or a replacement label provided by the expert. The members of rejected clusters are input to the next iteration of the clustering. These instances get a further chance to get into a cluster of good quality. Figure 5 shows a screenshot of our "cluster quality monitor", which coordinates the iterative clustering process by proposing clusters to the expert, sorting out approved clusters for labelling and rejected clusters for the next iteration.

A cluster label proposed by our software consists of the very frequent concepts of the cluster. Such a label may be:

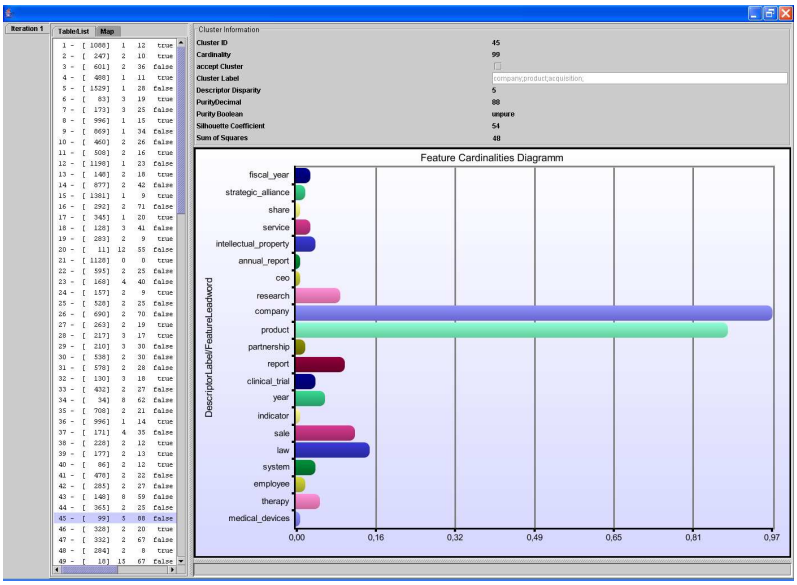

Figure 5: Cluster Quality Monitor

- a single concept from the input taxonomy or
- a combination of concepts appearing in the input taxonomy or
- a rephrasing of the proposed label, built by the domain expert

This label becomes the semantic tag annotating the text units in the cluster and extends the original taxonomy.

In our experiments, there was no involvement of domain experts. Rather, we have approved all clusters termed by our software as "good" and accepted the labels proposed for them. These clusters and labels, as produced during a total of 5 iterations, were delivered to the Biovista domain expert. The results of this evaluation by the human expert are not available yet, so that we present here two remarkable concept combinations that the Biovista partner has decided to disclose for the purpose of this publication.

An interesting concept combination is the proposed pair of concepts "approval" and "FDA"[4]: It is indicatory that the concept "FDA approval" may be an appropriate new concept for the taxonomy of the ontology. A further three-concept combination proposed as label for a good cluster consists of "patent_action","lawsuit" and "date"; the sentences in the cluster about patent-related lawsuits. In that case, the domain expert may decide to extend the taxonomy by a new concept "patent lawsuit" or add a relation between "patent" and "lawsuit" to the ontology. Such concept combinations which are characteristic for the corpus cannot be found by template-based methods in a straightforward way, because they occur independent of the phrasal form and are subject to frequency constraints.

[4]Food and Drug Administration

6 Discussion

It is impossible to ignore that the situation of "ontology shortage" that has characterized the first part of the nineties is now largely over. Therefore the reader might question why setting up ontologies (semi-)automatically is still very useful today. This section aims at providing motivation and illustrate domains where this automatic extraction is particularly needed.

We have today at our disposal, e.g., an impressive "upper level" tool like Philip Martin's WebKB-2 (see Martin (2003) and `http://meganesia.int.gu.edu.au/~phmartin/WebKB/`). This knowledge base reposes on an ontology derived from the "noun" section of WordNet 7.1 - where the most striking linguistic aspects have been removed and a more conceptual bias has been introduced. This basic (and very large) core is integrated with conceptual elements coming from other upper-level ontologies like DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering, see Masolo et al. (2003)) or Sowa's conceptual graphs (Sowa, 1999). Other well-known, free upper-level ontological tools are, e.g., SUMO (Suggested Upper Merged Ontology), developed by Tecknowledge Inc., see `http://ontology.teknowledge.com/` and Pease et al. (2002), and OpenCyc.[5]

With respect now to the proposals describing particular domains ("lower level" ontologies), we are confronted today to a real flood of specialized ontologies installed on the Web. The DAML site - we recall here that DAML (DARPA Agent Markup Language, see Hendler & McGuinness (2000)) is, with OIL (Fensel et al., 2000), one of the components of OWL (Smith et al., 2003), the new language proposed as standard for the description of ontologies by the W3C - includes an impressive lists of more than 700 sites (of different value) where specialized ontologies can be found, see `http://www.daml.org/ontologies/keyword.html`. And this site is far from being exhaustive: for example, it doesn't mention the comprehensive Business Process Ontology developed by Jenz and Partner, see `http://www.jenzundpartner.de/Resources/RE_OSSOntOWL/re_ossontowl.htm`.

Faced with this situation, (semi-)automatic tools for the set up of ontologies have still a very important role to play. Even if all the "ready to use" ontologies mentioned above already offer some sort of "pre-polished" and stabilized conceptual material, and (sometimes) some valued suggestions about the associated conceptual relations, this knowledge must be often restructured and, mainly, augmented, in order to appropriately cover the needs of specific institutions developing specific projects in particular domains. Moreover, in spite of the profusion of ontologies included in the DAML list, a quantity of domains are still waiting for some form of, at least preliminary, normalization: look at, e.g., the paucity of ontological material concerning domains which are currently the focus of a great deal of research activities, like violence, pornography and racism[6]. Eventually, we must also mention the fact that there are so many ontological proposals on the market as there are overlapping XML DTDs and contradictory EDI standards for quite a number of domains: therefore, (semi-) automatic tools can help with respect to the challenge of selecting the right components from each proposals and bringing them together.

Acknowledgments

The Parmenides project is funded by the European Commission (contract No. IST-2001-39023) and by the Swiss Federal Office for Education and Science (BBW/ OFES). For a detailed description of the project please see `http:/www.crim.co.umist.ac.uk/parmenides`.

The Parmenides consortium consists of the following partners (with responsible persons): **Biovista (GR)** Andreas Persidis; **Ministry of Defence (GR)** Thomas Mavroudakis, Spiros Taraviras; **Neurosoft (GR)** Giorgos Orphanos; **Otto-von-Guericke Universität Magdeburg (D)** Myra Spiliopoulou; **Coordinator: UMIST (UK)** Babis Theodoulidis, William Black; **Unilever (NL)** Hilbert Bruins Slot, Chris van der Touw; **University of Geneva (CH)** Margaret King; **University of Zurich (CH)** Fabio Rinaldi; **Wordmap (UK)** Will Lowe.

[5] see `http://www.opencyc.org/`, where a pointer to the "Cyc 101 Tutorial" can also be found).

[6] see `http://e-msha.msh-paris.fr/Agora/Tableaux\%20de\%20bord/Euforbia/`

References

Black, W. J., L. Gilardoni, F. Rinaldi, & R. Dressel (1997). Integrated text categorisation and information extraction using pattern matching and linguistic processing. In *Proceedings of RIAO97*, Montreal, pp. 321–335.

Black, W. J., J. McNaught, A. Vasilakopoulos, K. Zervanou, B. Theodoulidis, & F. Rinaldi (2003). CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations. Technical Report TR-U4.3.1, Department of Computation, UMIST, Manchester. `http://www.co.umist.ac.uk/~wjb/parmenides/tr-u4.3.1.pdf`.

Buchheit, M., F. Donini, & A. Schaerf (1993). Decidable reasoning in terminological knowledge representation systems. *Journal of Artificial Intelligence Research 1*, 109–138.

Chinchor, N. (1997). MUC-7 Named Entity Task Definition, Version 3.5. `http://www.itl.nist.gov/iaui/894.02/related\_projects/muc/proceedings/n%e\_task.html`.

Cunningham, H. (1999). JAPE: a Java Annotation Patterns Engine. Research Memorandum CS–99–06, Department of Computer Science, University of Sheffield.

Fensel, D., I. Horrocks, F. V. Harmelen, S. Decker, M. Erdmann, & M. Klein (2000). Oil in a nutshell. In *Knowledge Acquisition, Modeling, and Management - Proceedings of the European Knowledge Acquisition Conference, EKAW'2000*. Springer-Verlag.

Frantzi, K. T. & S. Ananiadou (1999). The C/NC value domain inpedented method for multi-word term extraction. *Journal of Natural Language Processing 6*(3), 145–180.

GENIA (2003). Genia project home page. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia.

Graubitz, H., M. Spiliopoulou, & K. Winkler (2001). The DIAsDEM framework for converting domain-specific texts into XML documents with data mining techniques. In *Proceedings of the First IEEE International Conference on Data Mining*, San Jose, CA, USA, pp. 171–178.

Hendler, J. & D. McGuinness (2000). The darpa agent markup language. *IEEE Intelligent Systems 15*(6), 67–73.

Ingria, B. & J. Pustejovsky (2002). TimeML Specification 1.0 (internal version 3.0.9). http://www.cs.brandeis.edu/~jamesp/arda/time/documentation/TimeML-Draft3.0.9.html.

Kôiti, H. The GDA Tag Set. `http://www.i-content.org/GDA/tagset.html`.

Kushmerick, N., D. S. Weld, & R. B. Doorenbos (1997). Wrapper induction for information extraction. In *Intl. Joint Conference on Artificial Intelligence (IJCAI97)*, pp. 729–737.

Martin, P. (2003). Knowledge representation, sharing and retrieval on the web. In N. Zhong, J. Liu, & Y. Yao (Eds.), *Web Intelligence*. Springer-Verlag.

Masolo, C., S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, & L. Schneider (2003). Wonderweb deliverable d17 - the wonderweb library of foundational ontologies. Technical report, Manchester: WonderWeb Project, University of Manchester.

Orphanos, G. & C. Tsalidis (2003). Deliverable 2/2: System architecture and technical specification. Technical report, PARMENIDES. IST-2001-39023.

Pastra, K., D. Maynard, O. Hamza, H. Cunningham, & Y. Wilks (2002). How feasible is the reuse of grammars for Named Entity Recognition? In *Proceedings of the 3rd Conference on Language Resources and Evaluation (LREC)*, Canary Islands, Spain.

Pease, A., I. Niles, & J. Li (2002). The suggested upper merged ontology: A large ontology for the semantic web. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*. Edmonton, Canada, July 28-August 1, 2002.

Pustejovsky, J., R. Sauri, A. Setzer, R. Gaizauskas, & B. Ingria (2002, July). TimeML Annotation Guideline 1.00 (internal version 0.4.0). `http://www.cs.brandeis.edu/~jamesp/arda/time/documentation/TimeML-Draft%3.0.9.html`.

Rinaldi, F., J. Dowdall, M. Hess, J. Ellman, G. P. Zarri, A. Persidis, L. Bernard, & H. Karanikas (2003a). Multilayer Annotations in PARMENIDES. In *The K-CAP2003 workshop on "Knowledge Markup and Semantic Annotation"*.

Rinaldi, F., J. Dowdall, M. Hess, K. Kaljurand, A. Persidis, B. Theodoulidis, B. Black, J. McNaught, H. Karanikas, A. Vasilakopoulos, K. Zervanou, L. Bernard, G. P. Zarri, H. B. Slot, C. van der Touw, M. Daniel-King, N. Underwood, A. Lisowska, L. van der Plas, V. Sauron, M. Spiliopoulou, M. Brunzel, J. Ellman, G. Orphanos, T. Mavroudakis, & S. Taraviras (2003b). Parmenides: an opportunity for ISO TC37 SC4? In *The ACL-2003 workshop on Linguistic Annotation, July 2003, Sapporo, Japan.*

Rinaldi, F., K. Kaljurand, J. Dowdall, & M. Hess (2003c). Breaking the deadlock. In *ODBASE 2003 (International Conference on Ontologies, Databases and Applications of SEmantics) Catania, Italy.*, Lecture Notes in CS. Springer Verlag.

Smith, M., C. Welty, & D. McGuinness (2003). Owl web ontology language guide - w3c proposed recommendation. Technical report, W3C.

Sowa, J. (1999). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove (CA): Brooks Cole Publishing Co.

TEI Consortium (2003). The text encoding initiative. `http://www.tei-c.org/`.

Winkler, K. & M. Spiliopoulou (2002). Structuring domain-specific text archives by deriving a probabilistic XML DTD. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*, Helsinki, Finland, pp. 461–474.

Wonsever, D. & J.-L. Minel (2001). Contextual Rules for Text Analysis. *Lecture Notes in Computer Science 2004*, 509–523

Xu, F. & H.-U. Krieger (2003). Integrating Shallow and Deep NLP for Information Extraction. In *Proceedings of RANLP 2003*, Borovets, Bulgaria.

Zarri, G. (1998). Representation of temporal knowledge in events: The formalism, and its potential for legal narratives. *Information and Communications Technology Law - Special Issue on Models of Time, Action, and Situations 7*, 213–241.

Zarri, G. (2003). A conceptual model for representing narratives. In R. Jain, A. Abraham, C. Faucher, & B. van der Zwaag (Eds.), *Innovations in Knowledge Engineering*. Advanced Knowledge International, Adelaide (Aus.).